Detection of Unknown Confounders

by Bayesian Confirmatory Factor Analysis

Emil Kupek

Department of Public Health
Centre for Health Sciences
Federal University of Santa Catarina
88040-900 Florianopolis, Brazil

Copyright © 2013 Emil Kupek. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Artificial data with known covariance structure was used to directly test confounding hypothesis using Bayesian confirmatory factor analysis with small variance informative priors. The priors were derived from two extreme scenarios of no versus maximum confounding via exposure variables. Large (N=5000) and small (N=100) sample analyses were performed for both continuous and binary variable models.

Both models showed large biases for all parameters in all analyses except for the direct effect of unobserved confounding on the outcome. Multivariate regression analyses yielded severely biased estimates of the exposure variables' effects. This problem was to some extent attenuated in Bayesian confirmatory factor analysis but the bias magnitude still remained large for the parameters in question.

In conclusion, Bayesian confirmatory factor analysis was shown reasonably precise to estimate direct confounding effect on the outcome but less so when this effect was mediated via exposure variables in both linear and binary regression models. It may be used a viable method for identifying unknown confounding variables and their relationship with the observed variables in the model.

Keywords: confounding, confirmatory factor analysis, Bayesian statistics, bias, multivariate regression

1 Introduction

Much of the methodological work in modern epidemiology has been dedicated to detection of and control for confounding, with main approaches being matching, stratification, restriction, and multivariate adjustment [6]. However, all of these assume that the confounding variables are observed, i.e. known and available in the data to be analyzed. On the other hand, the impact of unknown or unmeasured confounders is scarcely dealt with statistically in the vast majority of applied epidemiological studies, despite growing interest in the simulation studies to estimate the effect of measurement errors [10,19] and omitted covariates (which are not necessarily confounders) on the outcomes of interest. Omitted covariates lead to underestimation of the effect of exposure or intervention when these are associated with disease outcome, as well as potentially large reduction in study power [4].

Mainstream analytical strategies for detecting unobserved influences on outcome are largely based on the analysis of regression residuals. The shape and magnitude of their distribution, as well as differences in their variance (heteroscedasticity) may suggest omitted predictors which affect some observations more than the others, as revealed by outliers and so-called influential cases. However, these residuals are model-based and therefore conditional on the model being true, which is precisely in question. Even the examination of post-estimation statistical indices readily available in major statistical packages, such as Cook's distance, change in regression coefficients after removing particular observations and leverage values, is very rarely reported in applied epidemiological publications.

Bayesian statistics has provided most of the recent statistical methods to deal with unknown confounding due to its facility to incorporate prior knowledge of the associations studied into regression analysis in terms of prior distributions and extensively simulate plausible scenarios [18]. Although sensitivity analysis, both within Bayesian and frequentist framework, has been advocated in epidemiology for more than a decade [10,19,8,17,13,15] its wider dissemination in epidemiological practice remains rather restricted.

In the last two decades, much of the popularity of hierarchical (multilevel) modeling in epidemiology has been due to its capacity to reduce confounding [10,19] by accounting for the sources of unobserved heterogeneity, typically referred to as random effects [5]. After conditioning on observed covariates, residual variation may still be attributed to an unobserved (latent) variable which can meaningfully explain heteroscedasticity of the model residuals and identify the clusters of observations whose within-cluster variation is significantly smaller than the between-cluster variation. However, the interpretation of such unobserved variable is a conjecture which needs to be verified by additional data. From this perspective, multilevel models can reduce the effect of both known and unknown confounding variables on the outcome of interest (e.g. site-specific

influences) but cannot make statistical inference about the covariance structure of confounding.

On the other hand, structural equation modeling (SEM) can make such inferences providing certain assumptions. The key idea behind this approach is that unknown confounding is by definition a latent variable which influences both observed exposure variables and the outcome of interest in a regression model. This leads to a specific hypothesis which can be tested by confirmatory factor analysis (CFA) as it is known in SEM terminology. The first part of CFA is a common factor analysis model, also known as “measurement model”, where a latent variable (factor score) is derived as a linear combination of the observed independent variables. The second part of CFA, referred to as “structural”, regresses the outcome on both latent and observed variables, including those which are part of the “measurement model” (Figure 1). CFA therefore provides standard regression estimates adjusted for the latent variable representing unknown confounding.

A major problem with above method is that unless certain assumptions about model parameters such as measurement errors, regression coefficients or latent variable variance are made, the model cannot be identified. Standard Bayesian approach faces the same problem [13]. However, Bayesian CFA provides an excellent framework for adding reasonable distributional assumptions in order to make the model identifiable. The aim of this paper is to demonstrate how to apply Bayesian CFA to detect unknown confounding in both linear and binary (logistic) regression.

In the following sections, confounding and related concepts are defined as distinct types of model misspecification first, followed by the description of the methodology and the data used, results and discussion of principal findings.

Although “model misspecification” is a rather general term for a variety of omitted model parameters, it is mainly referring to omitted independent (exposure) variables in regression models. The omitted variables can be observed or unobserved (latent) as depicted in Figure 1. It is the covariance structure of the variables fitted that makes the difference between statistical models dealing with this issue. In the simplest case, the omitted variable is present in the data collected and its association with the outcome just needs to be verified in a regression model (X3 in Figure 1a). When omitted variable is unobserved (Z in Figure 1a), a random effects or multilevel model can be used to estimate differential impact of this variable on the regression parameters of observed independent variables across subpopulations or even individuals.

Confounding is a specific case of omitted independent variable, either observed (X3 in Figure 1b) or unobserved (Z in Figure 1b), when this variable is regressed on both outcome and other independent variables in the model [12]. Propensity score model hypothesizes modifying effect of an unobserved variable (Z in Figure 1c) on exposure variable (X1 in Figure 1c) where the former is correlated with other independent variables in regression model. The latent variable Z is usually estimated by principal components analysis and subsequently

fitted as control variable or used for stratification in regression analysis. Unobserved heterogeneity such as selection bias in recruiting participants is sometimes explored in this way.

2 Methods

Artificial data with known covariance structure are necessary to evaluate competing methods, in this case multivariate regression as a standard method versus CFA as a novel method proposed for testing the confounding hypothesis. Both continuous and binary variables example are presented, each with a large and small sample size version. The data generating process was as follows. Confounding variable was generated as a random normal variable with mean zero and variance of one (denoted as $f1$), and so were four error components (denoted $e1$ - $e3$ for independent variables and $d1$ for the dependent variable). Following equations were used to generate continuous outcome (y) and three exposure variables ($x1$ - $x3$) in a regression model where all observed variables are influenced by unobserved latent variable (f) representing a confounder:

$$x1 = 0.2 f + 0.2 e1$$

$$x2 = 0.1 f + 0.2 e2$$

$$x3 = 0.3 f + 0.4 e3$$

$$y = 0.9 x + 0.2 x2 + 0.1 x3 + 0.8 f + d1$$

Binary variables were generated by logistic function via $\exp(x)/(1+\exp(x))$ with threshold 0.5 for the predicting variables and 0.7 for the outcome.

SEM representation of the above model uses circle for latent confounding and squares for observed variables, with dashed lines indicating the effect of confounding (Figure 2).

The model comparisons in terms of bias are presented in tables 1 and 2. For continuous variables, standard multivariate regression model was fitted first, followed by the latent variable regression only (the model with dashed lines only in Figure 2). These models are two extreme scenarios with none versus maximum confounding effect via exposure (X) variables in the regression model, respectively. The next model fitted assumed that the most likely confounding effect was the middle point (mean) of these extremes with the probability of confounding dropping exponentially with the distance from the middle point, thus leading to the Bayesian priors with above mean and normal distribution with standard deviation of 1/6 the difference between multiple regression and latent variable regression estimates for independent variables, the latter being zero by definition.

For binary variables, multivariate binary (logistic) regression was fitted first, followed by the latent variable only binary regression model. Priors for Bayesian SEM were derived using the same reasoning as for continuous variables, keeping in mind the logarithmic scale of the estimates. In order to provide all estimates on the same scale, the Bayesian CFA regression coefficients were standardized as

partial correlations (denoted r) and transformed to odds ratios (OR) via $(1+r)/(1-r)$. The latter is the inverse of the Yule' transformation which approximates Pearson's correlation based on OR via $(1-OR)/(1+OR)$. The above transformation of partial correlations yields estimates of adjusted OR from Bayesian CFA and their natural logarithms are directly comparable to the logit scale estimates from the other two models for binary variables. More details on Yule' transformation is available from other sources [1,9].

Bias was calculated as percent difference between estimated and true values. Confounding via exposure variables was estimated through indirect effects of the latent variable on the outcome mediated by the exposure variables in Bayesian CFA. These effects were subtracted from the corresponding direct effects of exposure variables on the outcome to correct the bias induced in this way. On the other hand, bias component due to the direct effect of confounding on the outcome was estimated by the latent variable regression coefficient. A small random sample of 100 observations was taken to illustrate the impact of sample size on bias.

Bayesian CFA used Gibbs sampling in Markov chain Monte Carlo (MCMC) algorithm to iteratively approximate posterior distributions of the model parameters [21]. Two parallel chains were run with convergence criterion of proportional scale reduction (PSR) factor close to 1 and the first half of the iteration treated as "burn-in" phase [2,3]. Posterior checking graphs with 100000 iterations were used to further verify the convergence process. Posterior predictive p-value was also used to evaluate the quality of the model obtained.

Statistical software Mplus 6.11 (Los Angeles, Muthén & Muthén) was used for all analyses [14]. Analytical computer code is available from the author upon request.

Results

Both continuous and binary models showed large biases for all parameters in all analyses except for the direct effect of unobserved confounding on the outcome (Tables 1 & 2). Small sample size ($N=100$) did not alter above finding. Latent variable only regression grossly overestimated the confounding effect on the outcome in all models. Both linear and binary multivariate regression analyses yielded severely biased estimates of the exposure variables' effects. This problem was to some extent attenuated in Bayesian CFA but the bias magnitude still remained large for the parameters in question.

Convergence of Bayesian models was checked and confirmed by proportional scale reduction being close to 1 [2] after 100000 iterations. Graphical checks of convergence included posterior parameter distributions and trace plots, autocorrelation plots, posterior predictive scatterplots and distribution plots. The latter showed 95% confidence interval (CI) for the difference between the observed and the replicated chi-square values from -14.630 to 15.589, with

posterior predictive p-value of 0.491, whereas corresponding figures for the small sample were the 95% CI from -15.398 to 14.886 ($p=0.496$). For the large sample binary variables model, the 95% CI was between -15.163 and 15.347 ($p=0.496$), while the small sample binary variables model yielded the 95% CI between -14.355 and 16.999 ($p=0.443$). All these analyses pointed to convergence and excellent posterior predictive qualities of the models analyzed.

Discussion

Bayesian CFA detected direct confounding effect for both linear and binary regression with reasonable precision. Indirect confounding effects mediated through the exposure variables in regression yielded larger biases but still considerably lower than in multivariate regression.

Bayesian CFA can be used for statistical inference of the hypothesis of no confounding via one-tailed t-test by dividing the parameter with its posterior standard deviation. However, large variance of the latent confounding variable considerably reduced the power of this test and yielded statistically non-significant results for binary variables (p-values of 0.231 and 0.151 for large and small sample, respectively). Large latent variable variance is likely to be the problem in other situations where no additional information is available on its spread from previous research. Informative priors may be used with real data to mitigate this difficulty, as opposed to the artificial data example.

Prior elicitation is of crucial importance in Bayesian CFA. The method applied here used logical bounds for two extreme scenarios with minimum and maximum confounding effect via independent variables in regression and assumed normal distribution over the interval between these two points. More complex covariance structures may require different methods for prior elicitation which are beyond the scope of this paper but two of them may be particularly useful in this context.

MacLehose [17] showed how linear programming can be used to determine the bounds of the causal effect on the basis of the observed variables and derive plausible confounding scenarios for standard sensitivity analysis. Other methods may be more suitable for gauging the effect of highly correlated exposure variables [16,7].

The major advantage of Bayesian CFA over other methods to detect the impact of unknown confounding variables (e.g. random effect models) is its capacity to test directly a pre-specified confounding hypothesis by virtue of theory-driven small-variance priors [14]. Such models cannot be identified by maximum likelihood estimation unless additional constraints are made. The number of such constraints equals the square of the number of latent variables [11] and these are rarely based on substantive theory or justified statistically. In addition, Bayesian posterior distribution does not assume multivariate normal distribution of model parameters and is therefore more robust in its inference,

particularly for small samples and for the models with many parameters relative to the sample size [14].

Several limitations of this work should be kept in mind. First, only a relatively simple covariance structure was examined here, leaving aside important issues such as Bayesian CFA performance for a wider range of models beyond the scope of this paper. Second, the influence of sample size and different variable types has not been explored systematically. Third, the impact of model noise (e.g. the ratio of model to error variance) is yet another important factor [7] left for future research. Finally, as the exposure prevalence decreases, so does the bias due to confounding [6], thus limiting its impact for rare exposure events. Varying exposure prevalence would have provided more realistic results but was not the objective of this work.

The strengths of this work include the knowledge of true model and therefore an exact bias measure for all models due to the artificial nature of the data analyzed, as well as both large and small sample evaluation. It is well known that prior distribution has larger influence on Bayesian estimates in small samples. In addition, large number of iterations and extensive convergence and model checking provided a solid basis for Bayesian CFA estimates.

Although the example presented here used a continuous latent variable for confounding, it can be easily extended to categorical confounding variables, leading to latent class analysis [14]. Unknown factors causing effect modification may be explored in this way. Propensity score models can be estimated in one step by Bayesian CFA instead of a two-step procedure with principal components analysis to derive the score which is subsequently used for statistical adjustment in regression. Missing values can be estimated by multiple imputation [20]. A variety of dependency structures, such as spatial, temporal (longitudinal) and survival time models can be analyzed by Bayesian SEM [14]. Once confounding variable has been estimated, other forms of statistical adjustment may be employed, most notably a stratified regression analysis.

In conclusion, Bayesian CFA was shown reasonably precise to estimate direct confounding effect on the outcome but less so when this effect was mediated via independent variables in both linear and binary regression models. It is a viable method for identifying unknown confounding variables and their relationship with the observed variables in the model.

References

- [1] A. Agresti, *An introduction to categorical data analysis*, John Wiley & Sons, New York, 1996.

- [2] A. Gelman and D.B. Rubin, Inference from iterative simulation using multiple sequences, *Statistical Science*, 7(1992), 457-511.
- [3] A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin, *Bayesian data analysis*, Second edition, Chapman & Hall, Boca Raton, 2004.
- [4] A. Negassa and J.A. Hanley, The effect of omitted covariates on confidence interval and study power in binary outcome analysis: a simulation study, *Contemporary Clinical Trials*, 28(2007), 242-248.
- [5] A. Skrondal and S. Rabe-Hesketh, *Multilevel and Longitudinal Modeling Using Stata*, College Station, TX, Stata Press, 2005.
- [6] B.M. Psaty, T.D. Koepsell, D. Lin, N.S. Weiss, D.S. Siscovick, F.R. Rosendaal, M. Pahor and C.D. Furberg, Assessment and control for confounding by indication in observational studies, *Journal of American Geriatric Society*, 47(1999), 749-754.
- [7] D.C. Thomas, J.S. Witte and S. Greenland, Dissecting effects of complex mixtures: who's afraid of informative priors?, *Epidemiology*, 18:(2007) 186-190.
- [8] D.Y. Lin, B.M. Psaty and R.A. Kronmal, Assessing the sensitivity of regression results to unmeasured confounders in observational studies, *Biometrics*, 54(1998), 948-963.
- [9] E. Kupek, Beyond logistic regression: structural equations modelling for binary variables and its application to investigating unobserved confounders, *BMC Medical Research Methods*, 6(2006), doi:10.1186/1471-2288-6-13.
- [10] J. Schwartz and B.A. Coull, Control for confounding in the presence of measurement error in hierarchical models, *Biostatistics*, 4(2003), 539-553.
- [11] K. Hayashi and G.A. Marcoulides, Examining identification issues in factor analysis. *Structural Equation Modeling*, 13(2006), 631-645.
- [12] K.J. Rothman KJ and S. Greenland, *Modern Epidemiology*. 2nd ed. Philadelphia: Lippincott Williams & Wilkins; 1998.
- [13] L.C. McCandless, P. Gustafson and A. Levy, Bayesian sensitivity analysis for unmeasured confounding in observational studies, *Statistics in Medicine*, 26(2007), 2331-2347.

- [14] L.K. Muthén and B.O. Muthén, *Mplus User's Guide*, Sixth Edition, Muthén & Muthén, Los Angeles, 2010.
- [15] O.A. Arah, Y. Chiba and S. Greenland, Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders, *Annals of Epidemiology*, 18(2008), 637-646.
- [16] R.F. MacLehose, D.B. Dunson, A.H. Herring and J.A. Hoppin, Bayesian methods for highly correlated exposure data, *Epidemiology*, 18:(2007), 199-207.
- [17] R.F. MacLehose, S. Kaufman, J.S. Kaufman and C. Poole, Bounding causal effects under uncontrolled confounding using counterfactuals, *Epidemiology*, 16(2005), 548-555.
- [18] S. Greenland, Multiple-bias modeling for analysis of observational data (with discussion), *Journal of the Royal Statistical Society (A)*, 168(2005), 267-308.
- [19] S. Greenland, Variable Selection versus Shrinkage in the Control of Multiple Confounders, *American Journal of Epidemiology*, 167(2008), 523-529.
- [20] S. Van Buuren, *Flexible imputation of missing data*, Chapman & Hall/CRC; Boca Raton, 2012.
- [21] T. Asparouhov and B.O. Muthén, Exploratory structural equation modeling. *Structural Equation Modeling*, 16(2009), 397-438.

Figures

a)

b)

Figure 1. Graphical representation of an unobserved confounding variable model. Observed and unobserved variables are represented in boxes and circles, respectively. Dashed arrows represent direct effects of the unobserved confounding variable Z . Outcome and exposure variables are denoted as Y and X , respectively. Error components are denoted as e with subscript.

Figure 2. Three common scenarios of regression model misspecification: a) omitting observed (X3) and unobserved (Z) exposure variables influencing the outcome of interest (Y), b) observed (X2) and unobserved (Z) confounding variables influencing both outcome and exposure, and c) propensity score model with unobserved (Z) variable being correlated with independent variables in the model and also influencing the outcome (Y) through an exposure variable (X1). Observed and unobserved variables are represented in boxes and circles, respectively. Dashed arrows represent direct effects of the unobserved variables.

Table 1. Bias for regression coefficients in multivariate linear regression (MLR), latent variable regression (LVR) and Bayesian confirmatory factor analysis (BCFA) for continuous variables

Sample size	Model	Regression coefficients (standard error)			
		X1	X2	X3	F1 ^b
5000	True	0.9	0.2	0.1	0.8
	MLR	2.338 (0.062)	0.911 (0.073)	0.605 (0.034)	0 ^a
	Bias (%)	160	355	505	-100
	LVR	0 ^a	0 ^a	0 ^a	5.432 (0.135)
	Bias (%)	-100	-100	-100	579
	BCFA	1.756 (0.208)	0.605 (0.125)	0.378 (0.088)	0.720 (0.246)
	Confounding F1→X _n →Y	0.729 (0.136)	0.133 (0.035)	0.247 (0.069)	0
	Corrected for confounding	1.056	0.405		0.720
100	Bias (%)	17	205	131	-10
	MLR	2.009 (0.490)	0.944 (0.588)	0.527 (0.262)	0 ^a
	Bias (%)	123	372	427	-100
	LVR	0 ^a	0 ^a	0 ^a	4.741 (0.952)
	Bias (%)	-100	-100	-100	493
	BCFA	1.609 (0.319)	0.665 (0.267)	0.532 (0.160)	0.818 (0.312)
	Confounding F1→X _n →Y	0.747 (0.221)	0.213 (0.122)	0.286 (0.101)	0
	Corrected for confounding	0.862	0.452	0.246	0.818
	Bias (%)	-4	226	246	2

^a model-fixed value

^b latent factor score (true factor loadings of 0.2, 0.1 and 0.3 for X1, X2 and X3, respectively) as unknown confounding variable

Table 2. Bias for regression coefficients in multivariate binary regression (MBR), latent variable regression (LVR) and Bayesian confirmatory factor analysis (BCFA) for binary variables

Sample size	Model	Regression coefficients (standard error)			
		X1	X2	X3	F1 ^b
	True (logit)	0.9	0.2	0.1	0.8
5000	MBR logit	1.552 (0.078)	0.744 (0.072)	0.903 (0.074)	0 ^a
	Bias (%)	72	272	803	-100
	LVR logit	0 ^a	0 ^a	0 ^a	1.351 (0.156)
	Bias (%)	-100	-100	-100	69
	BCFA linear ^c	0.269 (0.202)	0.106 (0.086)	0.082 (0.150)	0.385 (0.417)
	Confounding F1→X _n →Y	0.173 (0.120)	0.043 (0.034)	0.047 (0.087)	-
	Corrected for confounding	0.094	0.063	0.035	0.385
	BCFA OR ^d	1.207	1.134	1.072	2.252
	BCFA logit ^e	0.189	0.126	0.070	0.812
	Bias (%)	-79	-37	-30	1
100	MBR logit	0.672 (0.502)	0.391 (0.480)	0.736 (464)	0 ^a
	Bias (%)	-25	95	635	-100
	LVR logit	0 ^a	0 ^a	0 ^a	0.415 (0.460)
	Bias (%)	-100	-100	-100	-48
	BCFA linear ^c	0.086 (0.206)	0.110 (0.162)	0.236 (0.161)	0.580 (0.427)
	Confounding F1→X _n →Y	0.065 (0.144)	0.077 (0.108)	0.163 (0.106)	0
	Corrected for confounding	0.021	0.033	0.073	0.580
	BCFA OR ^d	1.043	1.068	1.157	3.762
	BCFA logit	0.042	0.066	0.146	1.325
	Bias (%)	-95	-67	46	66

^a model-fixed value

^b latent factor score (true factor loadings of 0.2, 0.1 and 0.3 for X1, X2 and X3, respectively) as unknown confounding variable

^c SEM partial correlation estimates

^d adjusted odds ratio estimates using $(1+r)/(1-r)$, the inverse of Yule's transformation

Received: April 16, 2013